Human Language Technologies Evaluation in the European Framework.

J. Mariani
P. Paroubek

Human Language Technologies Evaluation in the European Framework.

J. Mariani, P. Paroubek

LIMSI-CNRS
BP 133, 91403 Orsay Cedex (France)
[mariani/pap]@limsi.fr

Abstract

Several initiatives have been conducted in Europe on Human Language Technologies Evaluation, both for spoken and written language processing. Some have been supported within the programs of the European Commission, such as SQALE, DISC or EAGLES. Others have been conducted within national programs, such as the GRACE action at CNRS, or the Aupelf-Uref ARCs, or internationally (Senseval/Romanseval) . A proposal for installing a comparative technology-evaluation infrastructure in the Fifth Framework program has been discussed within the Telematics-LE ELSE project, and the conclusions have been submitted to the European Commission. The EU-US Multilingual Information Access and Management (MLIAM) common initiative may provide a possible framework for cooperative activities in this area.

Evaluation activities in Europe

Several projects regarding Language Technologies evaluation have been conducted in Europe. At the level of the European Commission, one may especially mention SQALE, on Speech recognition evaluation, DISC on the design of best practice for dialog system development, Eagles, on Language Engineering standards, with a working group addressing the evaluation topic, and ELSE, which focuses on defining a black-box technology evaluation infrastructure within the future EC framework programs. Most of those projects are of limited size, and of short term duration, corresponding to the duration of a FP. ELSE proposes to create a more permanent infrastructure to capitalize on the experience gained throughout the projects.

Aupelf-Uref Francil ARCs (Coordinated research Actions) and Grace national projects, as well as the Senseval/Romanseval international project, are closer to the competitive evaluation scheme that has been pursued in the US.

The FRANCIL ARCs
(J. Mariani, 1998)

The AUPELF-UREF (Association of Francophone Universities) decided to launch in June 1994 a research network on Language Engineering, called Francil (Francophone Network on Language Engineering) (J. Mariani & F. Néel, 1995), with J. Mariani as coordinator, assisted by F. Néel as deputy coordinator. The goal of the Francil network is to ensure a good relationship between laboratories working in the field of Language Engineering, for the processing of the French language, either spoken or written language. The total Budget is about 4 MEcu over 4 years (1 MEcu for the network, including Cooperative Research Actions (ARP), 1 MEcu for a training program and 2 MEcu for the Strategic Research Actions (ARCs), based on the evaluation paradigm).

The ARCs, funded by the Fonds Francophone de la Recherche, cover both written language processing and spoken language processing with a total of 7 evaluation topics. For each topic, there exist 3 tasks: i) Organize the test campaign (involving an organizer and a coordinating committee), ii) Provide data (raw or annotated), iii) Participate in the test campaigns.

A Call for offer was sent on July 1994, with a deadline on November 1, 1994. 89 proposals were submitted. In March 1995, 50 proposals were selected, including 35 laboratories from 4 countries (Belgium, Canada, France, Switzerland). An evaluation campaign is conducted every two years. The first one took place in 1996-1997. The next one covers the period 1998-2000. Workshops were organized for each action, as satellite events of the Aupelf JST'97 conference, organized by Francil in April 1997.

In the domain of « Written language resources and systems evaluation » (ILEC), 4 actions have been initialized :

- A1 Natural Language access to textual information
- A2 (Bi/Multi)lingual corpus alignment
- A3 Automated terminological database design
- A4 Message understanding

In the domain of « Spoken language resources and systems evaluation » (ILOR), 3 actions have been initialized :

- B1 Voice dictation
- B2 Vocal dialog
- B3 Text-to-Speech synthesis

The organizers are, for Written Language Processing, A. Coret, now L. Schmidt (INIST (F), A1), J. Véronis (LPL (F), A2), C. Jouis and W. Mustafa (Univ. of Lille (F), A3) and P. Sabatier (LIM (F), A4). For Spoken Language Processing, J.M. Dolmazon, now P. Escudier (ICP (F)), B1 & B2) and A. Marchal, now B. Teston (LPL (F), B3). An international Advisory Committee comprising 6 members for ILEC and the same for ILOR participates in the selection of the proposals and in the evaluation of the program every year.

Written language processing (ILEC)

ARC A1 : Text Retrieval
(A. Coret et al., 1997a, A. Coret et al., 1997b)

8 laboratories participated in this action. The data consists of 3 different corpora : i) a corpus of the « Le Monde » newspaper, including, for training the systems, 15,000 documents and 11 topics (extended questions) and, for testing, 15,000 documents and 15 topics ; ii) a corpus of INIST scientific abstracts (extracted from the Francis and Pascal databases without domain restriction), including, for training the systems, 150,000 documents and 15 topics, and, for testing, also 150,000 documents and 15 topics ; iii) a corpus of books on the ethnology of Melanesia (about 50 books). Unfortunately, the agreement requested from the editors was not obtained quickly enough to consider this last corpus for the first campaign. The evaluation metrics consist of the Precision-Recall measures (% of documents retrieved which are correct vs % of correct documents which have been retrieved). The dry run and the test were completed for the first campaign, which was considered as an exploratory phase. The second campaign started in September 1998. Part of the evaluations are conducted over the Internet.

ARC A2 : Text alignment
(J. Véronis, 1997)

The task here is the alignment between the same texts written in French and English. 6 laboratories participated in the action (CITI (Canada), CRIN, LIA, IDL (France), ISSCO (Switzerland), UCREL Lancaster (UK)). In the first campaign, it was decided to consider sentences as the units to be aligned. The corpus comprises different types of texts: excerpts of the Official Journal of the European Union (JOC) (provided through the EU Multext project, 10 MWords in total / 1.2 Mwords per language) and CCITT technical texts (provided by the EU Crater project, 3 Mwords in total / 1 Mwords per language), provided by LPL, the BAF (« Bitextes Anglais-Français », 400 KWords for each language) provided by CITI, and fiction texts (« Le Désert des Tartares » and « Le Petit Prince »), provided by CRIN. The results are given as Precision-Recall measure (% of alignments produced which are correct vs % of source sentences correctly aligned, considering words or characters). The tests have been conducted in November/December 1997. Word alignment has been considered in the second campaign which started in September 1998. First results were reported at a workshop common with the Senseval and Romanseval actions (see further).

ARC A3 : Terminology extraction from texts
(A. Béguin et al., 1997, C. Jouis et al., 1997)

This action has 8 participants. The corpus consists of the SPIRALE Journal (Research in Education), including 19 issues of about 200 pages each, which have been manually indexed by experts and for which there exist a thesaurus and a list of key-words. Two other corpora are considered in the second campaign (from the Renault car company and the INRA Agriculture Research Agency). The different systems which are tested have different functionalities and provide different outputs ((ordered) terms, grammatical network, semantic graphs...). The evaluation is presently qualitative, and is provided by experts on the basis of the analysis of the usability of the information provided by the systems.

ARC A4 : Message Understanding
(P. Sabatier et al., 1997)

The result of the Call for Proposal on this topic was not sufficiently large to launch a complete action. Given the importance of the field, it was decided to install however for two years a Working Group, including 3 laboratories. This Working Group has produced a final report in November 1997, where it compiles a list of systems, cluster them into different categories, and propose to use the DQR (Documents-Questions-Responses) protocol (J.Y. Antoine et al. 1998) to assess them.

Spoken Language Processing (ILOR)

ARC B1 : Voice dictation
(J.M. Dolmazon & al., 1997, M. El Bèze & al., 1997)

The task consists in newspaper text dictation. The « Le Monde » newspaper has been chosen. 5 laboratories (CRIM, INRS (Canada), CRIN, Laforia, LIMSI (France)) participated in this action and 10 large vocabulary continuous speech recognition systems have been tested, gathered in 3 different test conditions (i) 20 KW, ii) 64 KW and iii) unlimited size vocabulary). The BREF speech corpus, designed at Limsi, has been distributed for training the systems, either as the BREF-80 subset (1 CDROM (80 speakers pronouncing 5000 sentences)), or as the full BREF corpus (12 CDROM (120 speakers pronouncing all material (100 hours))). A written language corpus, also provided by Limsi, has been distributed for training the language models. It consists in two years of the « Le Monde » newspaper (1987-1988, 40 MWords). A common lexicon (BDLex) was provided by IRIT, including the phoneme transcriptions, together with the list of the most frequent 20 Kwords and 64 KWords in the language corpus, and 4 Language Models (LM) (Bigram / Trigram, 20 KWords / 64 KWords). The test conditions were constrained for categories i) and ii) : the systems should use the 20 KWords (resp. 64 KWords) list and should be trained using the BREF corpus. The use of the provided Language Models was not mandatory, but the Le Monde data used for training should be anterior to May 1996 (Dry Run) or November 1996 (Test). The test data consists of two sets : T, a 600 sentences corpus, with open Out-Of-Vocabulary (OOV) word rate, and T', a 300 sentences corpus, with controlled OOV rate (less than 3%), as a subset of T. The « Dry Run » data (2 hours) has been built by 20 Speakers (12 male, 8 female). The prompts were given (Le Monde, May 1996). The test data (2 hours) has been built by 20 speakers (10 male, 10 female). In that case also, the prompts were given (Le Monde, November 1996). The results were returned in March 1997 and computed using the NIST/Sclite V3.0 software. In the adjudication phase, 474 claims were made

by 3 participants, and 94% of those claims were accepted. The description of the systems was provided by each participant. Results were reported as general word recognition rates, and the influence of various parameters was studied (speaker, speaking rate, male vs female etc), for each system and overall.

Within B1, a specific sub-action is conducted on the testing of Language Models. Several measures and protocols have been considered : i) computation of the perplexity, for missing word prediction : the systems bet on what may be a missing word, ii) testing various Language Models on the same recognized word lattice, iii) evaluate the Language Models as add-ons to an existing acoustic speech recognition system.

The content of the second B1 test campaign is presently being discussed. There is a possible extension to dialectal / regional variants of the French language, and to more challenging tasks, such as speech in noisy conditions and Broadcast News transcriptions, while retaining the former Le Monde dictation task in order to assess the progress achieved since the first campaign.

ARC B2 : Vocal dialog

(J.Y. Antoine et al., 1997, J. Caelen et al., 1997, J. Zeiliger et al., 1997, S. Rosset et al., 1997)

This action aims at the evaluation of spoken language understanding and dialog systems. 5 laboratories participate in the action. A first step was to choose the task domain (providing tourist information). A second step was to produce dialog corpora. Two corpora have been designed : a Human - Human « Pilot » corpus, consisting in the recordings of dialogs at a Tourist Office in Grenoble (15 hours), conducted by CLIPS and a Human - Machine corpus, consisting in the recordings of actual dialogs, based on scenarios, with a voice dialog system, which provides tourist information in a train station. This is conducted by Limsi, in cooperation with the SNCF (French railroad company).

A lot of discussions took place on the evaluation metrics, as it appears difficult to define evaluation measures, or even on the protocols in the area of dialog systems evaluation. Several evaluation metrics may be considered (evaluation of the components (recognition, parsing, dialog handling, generation, synthesis, etc), evaluation of the dialog duration, of the number of turns, of the comfort and satisfaction of users, etc). The DQR (Documents, Question, Response) approach (J.Y. Antoine et al. 1998), proposed in the ARCs A4 action and PARADISE (Walker 1997) have also been considered within B2.

ARC B3 : Text-To-Speech synthesis

(F. Yvon et al., 1998)

The task here is to evaluate Text-to-Speech systems in French. 9 participants (Limsi, LIA, ENS Telecom, ICP (France), LAIP (Lausanne), LATL (Geneva) (Switzerland), K.U. Leuven, TCTS Mons (Belgium), INRS (Canada)) are present in this action, and 7 systems have been evaluated in the first campaign. 4 kinds of tests are considered here : i) Grapheme-to-phoneme conversion, ii) Prosody iii) Encoding (voice quality) and iv) evaluation of the complete systems.

In the first campaign, only grapheme-to-phoneme translation modules were evaluated. The first step was to agree on a common phonetic alphabet (one close to SAMPA (Wells, 1997), designed in the SAM EU project (Fourcin and Dolmazon, 1991) was chosen). A Dry Run took place in April 1997, on the « Le mot et l'idée » text (a basic French text, including 99 sentences). The NIST scoring, initially designed for evaluating speech recognition systems, was used here for aligning the reference corpus and each transcription coming from the different systems, and for detecting and counting the transcription errors. The error rates were comprised between 0% and 5.3 %. An adjudication phase took place and the test campaign was conducted in September 1997, on the « Le Monde » Newspaper (2,000 sentences, totaling 26,000 words). The phoneme error rates range from 0.5% to 7%, while the sentence error rate goes from 10% to 80%.

For the future, it is foreseen to test complete TTS systems and to consider several tasks (newspaper reading, inverse directory, Human-machine dialogs, Email reading...). Subjective evaluation tests will be conducted at LPL (Aix-en-Provence). A possible extension to dialectal / regional variants of the French language is also considered here.

The CNRS CCIIL « GRACE » Action

(J. Mariani et al., 1997a, G. Adda et al., 1998)

This action, sponsored by CNRS within the « Cognition, Intelligent Communication and Language Engineering » action (CCIIL), aims at evaluating morphosyntactic taggers for French. Two corpora have been made available for training in two domains : « Le Monde » newspaper (1989-1990) and the INALF Frantext corpus (French literature of the 19th and 20th centuries). Testing was conducted on embedded text (20,000 Words embedded in a larger 300 KWords corpus), for both types of domains.

Following a Call for participation, 20 laboratories (from France, Germany, Switzerland, Canada and USA) responded and 18 participated in the action (13 to the end). The EAGLES / EU-LE-Multext tag set was chosen as reference. Each participant used his tag set and provided a translation table or a translation process between its own tag set and the reference one. The results were computed as a combined measure of Precision (% of correct tags) and Decision (% of complete disambiguations) for 3 different conditions: i) comparison with the proprietary tag set, ii) with the reference tag set, and iii) within a class of systems. A dry run phase was completed and its results were discussed at a satellite workshop organized during the JST'97. The tests have been completed in April 1998 and the results for condition ii) have been available on the WEB since November 98

GRACE yielded a very interesting by-product, a tagged corpus in French of 1 MWords obtained from the data tagged by 13 participants. Hand-made corrections of those tags will result in making available at a low cost (only

around 10% of the data need to be hand validated) a large tagged reference corpus for the development and evaluation of morphosyntactic tagging in French. Such reference tagged text is also of interest for development and testing in other domains, such as grapheme-to-phoneme transcription for text-to-speech systems and machine learning through system combination.

The SENSEVAL/ROMANSEVAL pilot project
(A. Kilgariff 1998)

SENSEVAL is a pilot evaluation campaign for Word Sense Disambiguating systems working on English. It was run in collaboration with the ROMANSEVAL evaluation campaign, the same exercise applied to the French and Italian. SENSEVAL/ROMANSEVAL ran over 8 months, starting from December 1997. The dry run data samples were distributed in March 1998 and test training data were distributed in June 1998. The tests were done on the instances of 20 nouns, 20 adjectives and 20 verbs. Initially, about 35 teams claimed their interest in participating in SENSEVAL, and, in the end, the results of the evaluation of 21 systems (including the derived versions) were presented at the final workshop.

The SQALE project
(S.J. Young et al. 1997)

The SQALE project ran from 1993 to 1995. The topic was the comparative evaluation of 3 different speech recognition systems, for different languages. It involved TNO-IZF (The Netherlands) as organizer, Philips (Germany), University of Cambridge (UK) and Limsi-CNRS as participants, each with its own system. Cambridge tested two systems (HTK and Cu-Con). The task was the dictation of newspaper texts in: American English (the common language that had to be addressed by all participants), British English, French and German. The Wall Street Journal was used for English, Le Monde for French and the Frankfurter Rundschau for German. Each participant had to evaluate its system on the common language, and, at least, on another language. The participants finally did the tests on between 2 to 4 languages. One of the findings in the project is that if a system is better on the common language than another system, it will also be better on its own language. However, it may not be better on other languages. This may reflect the language-independent algorithmic aspects of the systems vs the language specific knowledge which has to be taken into account. Comparison with human performance was also studied.

The DISC project
(N.O. Bernsen et al. 1997)

Although there already exists one for software development, today, no reference methodology exists for the development of spoken language dialog systems. DISC (ESPRIT long term research concerted action), aims at identifying what constitutes the best practice in spoken language dialog systems development and evaluation. DISC proposes a reference model built from two viewpoints, first, the components of the systems which have been identified as representative of the latest developments in the field, second the procedures and methods which are considered to produce the best results by the main actors of the domain. DISC produced a set of guidelines and heuristics to help in determining how a given system relates with respects to the proposed reference model.

Two schemata were used to structure the work: first, a grid of properties and aspects specific to the development and the working of the various modules composing a dialog system (dialog management, speech recognition, etc.), second, a life cycle model for dialog system development, inspired from those used in software engineering.

Results of DISC will be further enhanced (packaging, search and access) and thoroughly tested for usability in the course of the follow-up project: DISC-2 which started in January 1999.

The Eagles Evaluation
(M. King et al. 1996)

EAGLES was launched in 1993 to define standards for certain aspects of language processing technology, among them: evaluation. In order to reach an audience as broad as possible, the EAGLES evaluation working group used as starting point the ISO 9000 norm series (in particular ISO 9126 for software) and proposes a methodology which is strongly user-oriented (it advocates the use of the consumer report paradigm). The methodology can be applied either for adequacy evaluation or progress evaluation (with emphasis on the former). To support its methodology EAGLES has developed a formalism (based on feature structures) for classifying products and users. In the course of the project and of TEMAA (an associated LRE project) case studies were performed on spelling checkers, grammar checkers and translators' aids. EAGLES-II (1995-1998) was a follow-up project, whose goals were to consolidate, extend and disseminate the results of EAGLES.

The ELSE project

ELSE, a 16 month LE-4 preparatory action which started in January 1998, aims to draw up a blueprint for an evaluation infrastructure which could be deployed in the scope of the IST Key Actions of the fifth framework program of the European Community. ELSE has identified the main differences between the deployment conditions of the paradigm in the United-States and in Europe and studied the potential benefit of deploying the evaluation paradigm (faster progress, acceleration of innovation transfer from research to development, contribution to building a clearer view of a field, increase of the amount of data, knowledge and tools available).

Among the five different possible approaches towards evaluation it has identified, ELSE has focused its efforts on Technology Evaluation and has described how Technology Evaluation is positioned with respect to the

other types of evaluation, in particular showing how it precedes and complements Usage-Oriented Evaluation in the development lifecycle of an application.

ELSE also found out that there is a need for a permanent infrastructure in Europe in order to capitalize across Framework Program boundaries on the experience gained throughout the evaluation projects and made suggestions on how to integrate evaluation in the Call for Proposals To this end, a list of candidate control tasks which are good candidates for starting a series of evaluation campaigns in a pro-active scheme has been identified. It could also be replaced by a single task of interest for both the speech and NL communities, such as crosslingual News on Demand systems (of the Informedia type (Wactlar et al., 1999)) evaluation, for example. As a fall-back position, the possibility to deploy evaluation in a re-active scheme as an add-on activity to technology based project clusters (grouping project according to the technology used or developed) has also been investigated.

ELSE proposes to address the multilingual aspects of deploying the evaluation paradigm in Europe either with cross lingual requirements or with the use of a common pivot language. ELSE made a first estimate of the cost of deploying evaluation on a large scale using. To complete the picture, a preliminary study of the legal issues raised by the deployment of evaluation was conducted. The results of ELSE will be available in Spring '99.

Towards FP5

The 5th Framework Program has just been launched. This 4-year program comprises 4 different Thematic Programs, one of which being Information Society Technologies (IST). This program will have a budget of 3.6 BEuro. Within IST, a Key Actions is devoted to Multimedia Content and Tools, which includes a specific action on Human Language Technologies, which should have a budget of about 125 MEuro. Human Language Technologies is organized in 3 Action Lines: Multilinguality in digital contents and services, Natural interactivity and Cross-lingual information management. The third action line will be launched in 2000, while the first two ones should be launched in March 1999. Apart from those three action lines, several Support Measures appear, including project clustering, networking, working groups, studies and dissemination and awareness actions accompany the IST program. Activities such as "Best practice actions", "Assessments", "Measuring user acceptance", "Assessing system performance", "Provision of suitable measurement and testing facilities" appear in the document. It may therefore allow for launching EU-wide common activities in this European exercise.

MLIAM

A cooperation agreement has been signed by the EU and the US recently. Within this agreement, an action on "Multilingual Information Access and Management" (MLIAM) addresses the topic of transatlantic cooperation between the NSF and the EC, based on the needs for globalization, for multilingual Information and for Cross-cultural communication. Several topics are contained in the Call for Proposal which was recently issued. It includes Research and Technology development, Standards for Linguistic Data, Multilingual ontologies, Linguistic Data Centers and preparatory actions. The funding mechanism is that, once the project has been selected, the partners are supported by their respective agencies. This may allow to support a joint EU-US effort of Human Language Technologies evaluation.

Conclusions

As a conclusion, we shall stress the importance of the evaluation paradigm in Language Engineering (both for Spoken and Written Language processing) (R. Cole, 1997, L. Hirschman & H.S. Thompson, 1997, J. Mariani, 1997b). It induces the necessity of defining precise evaluation metrics and the availability of well documented language resources for training and testing, produced in due time and in conformity with the specifications. It allows for a better understanding of the advantages and drawbacks of the different systems, approaches and methods, which are discussed during the workshops in the light of the test results which concern the same data, each participant trying to do its best on that task. Furthermore, we think that it is essential to take into account multilinguality when deploying the evaluation paradigm. because it is an essential characteristic of the information society in which we live. In the US, the evaluation paradigm has been used within the DARPA and NIST actions and programs, and reported since 1987, mostly on American English. It has been recently extended to other languages (Multilingual TREC). A proposal for preparing a possible Human Language Technologies Evaluation infrastructure within the EC 5th Framework program has been investigated within the EC FP4-Telematics ELSE project, and may allow for the introduction of evaluation in FP5. It is important to consider in this perspective the possibility of EU-US cooperation within the MLIAM program. as well as the coordination of national (e.g. French) or language specific actions (such as the Aupelf-Uref ones) with the EC effort in that area.

References

G. Adda, J. Lecomte, J. Mariani, P. Paroubek, M. Rajman, «The GRACE French Part-of-Speech Tagging Evaluation Task.», proceedings of the First International Conference on Language Resources and Evaluation (LREC), Granada, May 1998.

F. Yvon, P. Boula de Mareuil, C. d'Alessandro, V. Auberge, M. Bagein, G. Bailly, F. Bechet, S. Foukia, J.F. Goldman, E. Keller, D. O'Shaughnessy, V. Pagel, F. Sannier, J. Véronis, B. Zellner « Objective evaluation of grapheme to phoneme conversion for text-to-speech synthesis in French. », Computer Speech and Language, Vol. 12, pp393-410, October 1998

J Y Antoine, J Zeiliger, J Caelen, « RQA methodology: towards a qualitative evaluation of speech understanding and spoken dialog systems. », *Speech and Language Technology (SALT) Club Workshop on Evaluation in Speech and Language Technology*, June 17-18, 1997, University of Sheffield (UK)

J Y Antoine, J Zeiliger, J Caelen, « DQR Test Suites for a Qualitative Evaluation of Spoken Dialog Systems: from Speech Understanding to Dialog Strategy. », proceedings of the First International Conference on Language Resources and Evaluation (LREC), Granada, May 1998.

A. Beguin, C. Jouis, W. Mustafa, « Evaluation d'outils d'aide à l'extraction et à la construction automatiques de termes et de relations sémantiques », *JST Francil 1997 proceedings*, Aupelf-Uref, April 15-16, 1997, University of Avignon (F)

N.O. Bernsen, L. Dybkjaer, R. Carlson, L.Chase, N. Dahlback, K. Failenschmid, U. Heid, P. Heriterkamp, A. Jonsson, H. Kamp, I. Karlson, J. V. Kuppevelt, L. Lamel, P. Paroubek, «The DISC approach to Spoken Lanuguage System Development and Evaluation», proceedings of the First International Conference on Language Resources and Evaluation (LREC), Granada, May 1998.

J. Caelen, J. Zeiliger, M. Bessac, J. Siroux, G. Perennou, « Les corpus pour l'évaluation du dialogue homme-machine. », *JST Francil 1997 proceedings*, Aupelf-Uref, April 15-16, 1997, University of Avignon (F)

R. Cole, chapter editor, « Chapter 12 : Language Ressources », in R. Cole, J. Mariani, H. Uszkoreit, G.B. Varile, A. Zaenen, A. Zampolli, V. Zue eds, *« Survey of the State-of-the-Art in Human Language Technology »*, Giardini, Editori, 1997

A. Coret, P. Kremer, B. Landi, D. Schibler, L. Schmitt, N. Viscogliosi, « Accès à l'information textuelle en français : Le cycle exploratoire Amaryllis. », *JST Francil 1997 proceedings*, Aupelf-Uref, April 15-16, 1997a, University of Avignon (F)

A Coret, P Kremer, B Landi, D Schibler, L Schmitt, « Towards a methodology for evaluating information retrieval systems adapted to textual documents in the French language: the Amaryllis exploratory cycle », *Speech and Language Technology (SALT) Club Workshop on Evaluation in Speech and Language Technology*, June 17-18, 1997b, University of Sheffield (UK)

J.-M. Dolmazon, F. Bimbot, G. Adda, Marc El Bèze, J. C. Caerou,. J. Zeiliger, Martine Adda-Decker, « Organisation de la première campagne AUPELF pour l'évaluation des systèmes de dictée vocale. » , *JST Francil 1997 proceedings, Aupelf-Uref*, April 15-16, 1997, University of Avignon (F)

M. El Bèze, M. Jardino, F. Bimbot, « Une approche alternative pour le calcul de la perplexité. », *JST Francil 1997 proceedings*, Aupelf-Uref, April 15-16, 1997, University of Avignon (F)

A.J. Fourcin, J.M. Dolmazon, "Speech knowledge, standards and assessment", Proceedings of the XII International Congress of Phonetic Sciences, Aix-en-Provence, Vol. 5, 430-433, 1991

C. Jouis, W Elhadi, « AUPELF Project: Term and Semantic Relation Extraction Tools: Evaluation Paradigms », *Speech and Language Technology (SALT) Club Workshop on Evaluation in Speech and Language Technology*, June 17-18, 1997, University of Sheffield (UK)

Adam Kilgarriff, «SENSEVAL: An Excercise in Evaluating Word Sense Disambiguation Programs »,, in Proceedings of the First International Conference on Language Resources and Evaluation (LREC), Granada, May 1998.

M. King, B. Maegaard, J. Schutz, L. des Tombes et al., «EAGLES – Evaluation of Natural Language Processing Systems », final report, EAG-EWG-PR.2, Ocotber 1996.

J. Mariani, F. Néel, «Aupelf-Uref Actions Towards Language Resources and Evaluation», *COCOSDA workshop, September 22, 1998, Madrid, Spain.*

J. Mariani, J. Lecomte, P Paroubek, M Rajman, G Adda, « Progress report on the GRACE evaluation program for French Part-Of-Speech taggers », *Speech and Language Technology (SALT) Club Workshop on Evaluation in Speech and Language Technology*, June 17-18, 1997a, University of Sheffield (UK)

J. Mariani, "Some evaluation-based language engineering actions for French", Computer Speech and Language, Vol. 12, pp 303-316, October 1998

S. Rosset, S. Bennacef, J. Gauvain, L. Devillers, L. Lamel, « Corpus oral de renseignements touristiques. », *JST Francil 1997 proceedings*, Aupelf-Uref, April 15-16, 1997, University of Avignon (F)

P. Sabatier, P. Blache, J. Guizol, F. Levy, A. Nazarenko, S. N'Guema, R. Pasero, M. Rolbert, « Evaluer des Systèmes de Compréhension de Textes. », *JST Francil 1997 proceedings*, Aupelf-Uref, April 15-16, 1997, University of Avignon (F)

J. Véronis, « Une action d'évaluation des systèmes d'alignement de textes multilingues. », *JST Francil 1997 proceedings*, Aupelf-Uref, April 15-16, 1997, University of Avignon (F).

H.W. Wactlar, M.G. Christel, Y. Gong, A.G. Hauptmann, "Lessons learned from building a Terabyte Digital video library", IEEE Computer, pages 66-73, February 1999

M. Walker, D. Litman, C. Kamm, A. Abella, « PARADISE: A Framework for Evaluating Spoken Dialogue Agents. », proceedings of the 35th Annual Meeting of the Association of Computational Linguistics, ACL 97, 1997.

J.C. Wells, "The SAMPA Alphabet", web site: http://www.phon.ucl.ac.uk/home/sampa/home.htm, 1997

S.J. Young, M. Adda-Decker, X. Aubert, C. Dugast, J.L. Gauvain, D.J., Kershaw, L. Lamel, D.A. Leeuwen, D. Pye, H.J.M. Steeneken, A.J. Robinson, and P.C. Woodland. « Multilingual large vocabulary speech recognition: the european SQALE project. »,. Computer Speech and Language, 11(1):73-89, January 1997.

J. Zeiliger, J. Caelen, J.-Y. Antoine, « Vers une méthodologie d'évaluation qualitative des systèmes de compréhension et de dialogue oral homme-machine. », *JST Francil 1997 proceedings*, Aupelf-Uref, April 15-16, 1997, University of Avignon (F).